The Harder the Task, the Higher the Score: Findings of a Difficulty Bias

Hillary N Morgan
Drew University

Kurt W Rotthoff*
Seton Hall University
Stillman School of Business

Fall 2013

Abstract

Studies have found going first or last in a sequential order contest leads to a biased outcome; commonly called order bias (or primacy and recency). Studies have also found judges have a tendency to reward contestants they recognize with additional points, called reference bias. Controlling for known biases, we test for a new type of bias we refer to as 'difficulty bias', which reveals that athletes attempting more difficult routines receive higher execution scores, even when difficulty and execution are judged separately. Despite some identification challenges, we add to the literature by finding strong evidence of a difficulty bias in gymnastics. We also provide generalizations beyond athletics.

JEL: L10, L83, D81, J70, Z1

Keywords: Difficulty Bias; Sequential Order Judging; Judging Bias; Reference Bias

* Kurt Rotthoff at: Kurt.Rotthoff@shu.edu, Seton Hall University, JH 674, 400 South Orange Ave, South Orange, NJ 07079. Hillary Morgan: HillaryNMorgan@gmail.com. We would like to thank Angela Dills, Robert Tollison, Sean Mulholland, Rey Hernandez, Pete Groothuis, Ryan Rodenberg, Jay Emerson, Sarah Marks, the participants at the American Statistical Association's annual meetings and referees for helpful comments. Also a special thanks to the editor, Jeff Borland , for helping us clarify thoughts throughout the manuscript. Any mistakes are our own.

I. Introduction

Judgments are made in many areas of life: job interviews, refereed journal articles, marketing pitches, oral and written exam grades, auditions, sporting events, debates, or even stock analyst's estimates. In arenas where judges determine the outcome of an event, bias in the judging process can create problems. Biased judging potentially leads to questions about efficiency and fairness, particularly if it results in selecting less than optimal candidates (Page and Page 2010).

Judging and perception bias have been observed in a variety of situations. Psychologists show that sequential presentation of information can influence the way the information is processed (Mussweiler 2003). This idea has been carried over to other fields including economics (Neilson 1998; Sarafidis 2007; and Page and Page 2010) and marketing (Novemsky and Dhar 2005). Judging bias has been found in orchestra auditions (Goldin and Rouse 2000) and sequential voting through the "Idol" series (Page and Page 2010). Bias has also been found in basketball referees (Price and Wolfers 2010).

We test for bias in the judging of elite gymnastics. In particular, the gymnastics meet we analyze provides a uniquely suitable dataset: the order of competition is randomly assigned to a given country and the difficulty and execution of a routine are separately judged.[1] Following previous biases found in the literature, we control for performance order (primacy and recency) and reference bias. Despite some unit analysis challenges in our control for reference bias and identification issues concerning our lack of a perfect control for athlete ability, we add to the literature by finding strong evidence

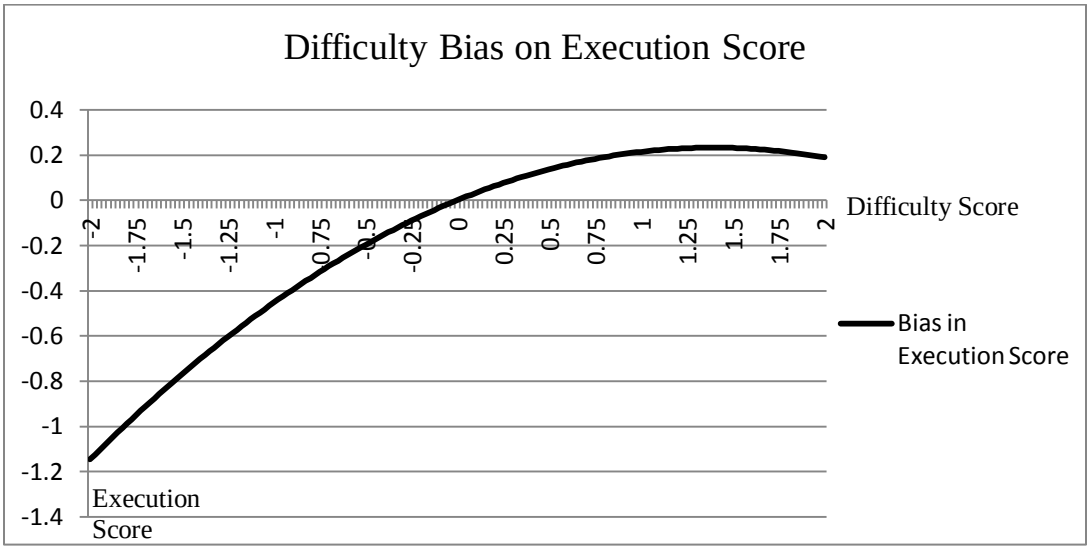
[1] Nearly random assignment of athletes in gymnastics is rare, making this a unique dataset. Separate panels for judging began in 2006. The event we use is the only elite level meet with numerous countries in attendance that meets both of these requirements at this point in time.

of difficulty bias; execution judges show a favorable bias for those athletes attempting more difficult routines.

Measuring difficulty bias requires data where judgment is delivered in two parts: difficulty and execution. This can be found in the world of elite level gymnastics. Elite gymnasts receive scores based on the difficulty of the task and the execution of this task. One panel of judges is charged to evaluate the execution, and only the execution, of the routine, with an independent panel of judges evaluating the difficulty, and only the difficulty, of each routine. In other words, execution judges should not be concerned with the difficulty of the routine and difficulty judges should not be influenced by the execution. Because the judges sit on separate panels, we can determine if the difficulty of the routine influences the execution score.

Using normalized data, mean zero and standard deviation of one, we regress execution score on difficulty score, with additional controls. We find that a participant's overall score is artificially inflated when that athlete attempts a more difficult routine. Figure I shows the extent of this bias. Increasing one's difficulty by one standard deviation artificially inflates the execution measure by 0.21 standard deviations.

Figure I: Impact Difficulty Bias on Execution Score.



Likewise, attempting a less difficult routine, one that is one standard deviation below the mean, decreases the execution score by 0.45 standard deviations.

This finding has major implications for the ability of judges to accurately rank individuals. In situations where judgment is passed on a given performance, participants may choose to execute a more difficult gymnastics routine, play a more complex piece of music at an audition, tackle a more challenging research topic when applying for a grant, or even use impracticable statistical approaches to impress a referee; all with the knowledge that the difficult act in question may influence the evaluator, resulting in a biased execution score.

The next section provides background on types of judging bias including order bias, reference bias, and others. We also outline the ways in which our dataset allows us to distinguish between forms of bias. Section three discusses an overview of the data and is followed in section four by the methodology, with the limitations of our data. Section five discusses our addition to the literature, difficulty bias, in detail. The last section concludes with policy implications.

II. Types of potential bias in sequential order events

The psychology literature looks at judgment bias in sequential order events, finding two key effects: a primacy effect and a recency effect. If primacy exists, the first person or people to perform are judged more accurately. If judges better remember late effects, a recency effect results. Gershberg and Shimamura (1994) and Burgess and Hitch (1999) conclude that in a sequential order contest it would be best to go either first or last, but not in the middle. The economics literature takes a different view on this idea. In situations where the scores of each contestant are finalized before the next contestant competes, as is the situation with our data, findings of an overall order bias are more common.

The overall order bias impacts a contestant's relative ranking depending on when in the event they compete. For example, Wilson (1977) finds evidence that the order of appearance in synchronized swimming influences the outcome. Analyzing artists that compete in "The Queen Elizabeth musical competition," Flôres and Ginsburgh (1996) find the day an artist competes impacts that artist's final standing. Bruine de Bruin (2005) studies both the "Eurovision" song contest as well as figure skating, finding those that perform later receive more favorable evaluations in both venues. Page and Page (2010) also find an overall order bias in the "Idol" song contest.

Damisch, Mussweiler, and Plessner (2006) find a sequential order bias, where one person's performance impacts the subsequent performer, in the 2004 Olympic Games. They find that a gymnast's score is influenced by the previous performance. However, there is no evidence of this type of bias in the 2009 World's meet, as found in Rotthoff (2013).We therefore focus on overall order bias.

The psychology literature also presents a 'reference bias' in judgment. People, or judges, may have a tendency to rate a participant relative to their expectations on that person's performance (Thibaut and Kelley 1959). In the workplace, raters who are more familiar with a worker tend to give more positive overall ratings than those that are not familiar with that individual (Kingstrom and Mainstone 1985). Tversky and Kahneman (1974) and Kahneman and Tversky (1996) describe heuristics, or the use of a representative tool, as a shortcut to process information. Findlay and Ste-Marie (2004) find that figure skating judges use this representative tool, in the form of athlete reputation, to judge a given athlete's performance, biasing the known athletes' scores upward.

In addition to order and reference biases, evidence of racial, gender, and nationalistic judgment biases have been discovered. For example, Glejser and Heyndels (2001) confirm the order bias results from Flôres and Ginsburgh (1996) concerning music competition and further find that women obtain lower scores in piano while contestants from the Soviet Union, prior to 1990, receive higher scores. Multiple other studies find a nationalistic bias in figure skating: Seltzer and Glass (1991) find a bias based on political loyalties, Sala, Scott, and Spriggs (2007) find a systematic bias based on the countries status as a "friend", "rival", or "enemy", and both Campbell and Galbraith (1996) and Zitzewitz (2006) find nationalistic biases. Emerson, Seltzer, and Lin (2009) find strong evidence of a nationalistic bias in Olympic diving and Segrest, Perrewe, Gillespie, Mayes, and Ferris (2006) find a negative ethnic bias in the hiring process. Racial bias is found by Price and Wolfers (2010) in professional basketball refereeing, by Parsons, Sulaeman, Yates, and Hamermesh (2011) in baseball as umpires call strikes, and by

Garicano, Palacios-Huerta, and Prendergast (2005) as referees favor the home team in soccer (football).

We hypothesize that when reference points are limited, judgment is made relative to a known element of the given task: Difficulty. Given the judges know what a difficult task is, they present biased scores when more difficulty exists.

III. Data

Gymnastics is uniquely able to distinguish the types of bias described in the previous section. We use data from the 2009 World Artistic Gymnastic Championships, held in London, England. Unlike the majority of large international gymnastics meets, this one only offered individual all-around and individual event competitions for male and female elite level gymnasts. This meet provides insight into the described forms of bias because there is no team competition.[2] More importantly, the meet randomly assigns each country one to three starting positions, based on the number of spots that country qualifies for. Each country's governing body then distributes the spots to their athletes.

Elite gymnastics also recently changed its scoring system, allowing us to separate the athletes' difficulty of performance from their execution. The difficulty and execution scores are awarded by separate panels of judges. The two scores are then added together and, after taking out any penalties, the final mark is awarded. Scores are given after each contestant, so each score is finalized before the next contestant makes their attempt. More detail on scoring is given later in this section.

[2] In team competitions the coach chooses athlete orders to optimize the team performance. This behavior removes the random performance order aspect that is valuable when conducting statistical analysis.

Gymnastics Basic Rules

In women's gymnastics there are four different events (vault, uneven bars, beam, and floor) while the men have six events (vault, floor, pommel horse, rings, high bar, and parallel bars). The structure of the competition allows for enough recovery time between events, so the athlete's performance on each event is independent. In the 2009 Worlds, each country could bring up to three athletes to compete in each event, but many athletes competed in multiple events at the meet. This is not unusual. Top talent is often good at multiple events and they compete for the all-around title, where their additive score for all individual events determines the winner. Based on their performance in the preliminary round, athletes can make finals in each individual event as well as for the all-around competition.

Most international competitions have a team competition built into each meet. Teams often strategically place their athletes to maximize the team score, which traditionally means ordering the athletes from the lowest expected score to the highest. This meet does not have this team aspect.

For each of the ten events, we observe between 106 and 134 performances; the number varies based upon the number of athletes attempting to make the finals in either the all-around or on a given event. Each event has a preliminary and final session, usually spaced a couple days apart. The finals are structured in a traditional gymnastics way, with the lowest scoring person going first. The goal in the preliminary round is to get the best spot in the finals competition and it is commonly known in the sport that the last spot is best. This aligns the incentives of the athletes; each athlete wants to perform their best in prelims in order to have the best position in the finals competition. For this reason we use

only preliminary scoring data and in this round their goal is always score maximization, thus the use of preliminary data does not bias the sample.

Gymnastics Scoring

In 2006 the gymnastics governing body, the FIG (Federation Internationale de Gymnastique), completely overhauled the scoring system for elite level gymnastics. This change came after an apparent judging controversy in the 2004 Athens Olympics. Scores are now divided into two parts: difficulty and execution, which sets this dataset apart from Damisch, Mussweiler, and Plessner (2006). The system now separates out the 'D' score, which is designed to exclusively measure the difficulty, and the 'E' score, which is designed to exclusively measure the execution score.

The difficulty score evaluates the content of the routine. Judges award points on three basic parts: the difficulty value of the routine, the demonstration of a fixed set of required skills, and added points for connecting certain elements.[3] On vault, the same difficulty score is awarded to every athlete that performs the same vault, as determined by the gymnastics Code of Points. On all other events, a panel of judges evaluates the difficulty score while the athlete performs. They then compare the score among themselves and post it. The difficulty score is theoretically infinite and is determined by the athlete because they decide what level routine to do, meaning it is exogenous to the judges.

The execution score evaluates how perfectly the athlete performs on that event. This score has a maximum value, and a starting value, of a 10.0 and salvages the part of the scoring system that made Nadia Comaneci a household name. From the beginning of

[3] An athlete's difficulty score can be increased when two elements are combined. The combination of elements is considered a more difficult task than doing them individually.

each routine, the judge takes away points for errors in form, execution, technique, artistry, and routine composition. The execution score is determined solely by the judges on the execution panel and will capture any bias in the judging process, if it exists.

The difficulty and execution scores are awarded by completely separate panels of judges. With the exception of vault, where the difficulty to be attempted is posted before the gymnast performs, the difficulty and execution scores are evaluated simultaneously and directly after the gymnast completes his or her routine.[4] The two scores are then added together, and after taking out any penalties (primarily given for athletes stepping out of bounds) the final mark is awarded. Scores are posted after each contestant, meaning each score is finalized before the next contestant makes their attempt. The average and standard deviation of scores for the 2009 World Gymnastics Championships are shown in table 1 (women) and 2 (men).[5]

Table 1 – Women's Events.

Summary Statistics (women)				
Variable	Vault	Uneven Bars	Balance Beam	Floor
Participants	107	113	118	113
Mean Difficulty Score	4.94	4.89	4.99	4.92
Standard Deviation of Difficulty Score	0.706	1.194	0.650	0.564
Mean Execution Score	8.24	6.91	7.21	7.37
Standard Deviation of Execution Score	0.904	1.517	1.161	0.778

[4] Although the vault number from the gymnastics Code of Points and implicitly the difficulty score for the vault is posted before the event, the athlete's difficulty rating can change if they complete a different vault than what has been posted.

[5] The mean and median are close, showing that any outliers are not driving the data.

Table 2 – Men's Events.

Summary Statistics (men)						
Variable	Parallel Bars	High Bar	Rings	Floor	Vault	Pommel Horse
Participants	127	127	126	134	122	132
Mean Difficulty Score	5.31	5.31	5.43	5.51	5.31	5.14
Standard Deviation of Difficulty Score	0.88	1.00	0.91	0.79	0.88	0.90
Mean Execution Score	8.07	7.80	7.94	8.16	8.07	7.68
Standard Deviation of Execution Score	0.78	0.85	0.66	0.96	0.78	1.17

Normalization

Because there is only one athlete that goes first and one that goes last on each event over the entire day of preliminary competition, we aggregate each of the ten events together and use the overall order of each event. Aggregation allows more observations and increases the validity of the estimates. However, because the mean and standard deviations are different on each event, we first normalize all men's and women's events to have a mean zero and a standard deviation of one; then aggregate the data together. The summary statistics for all aggregated events are in table 3.

Table 3 – Normalized data for All Events.

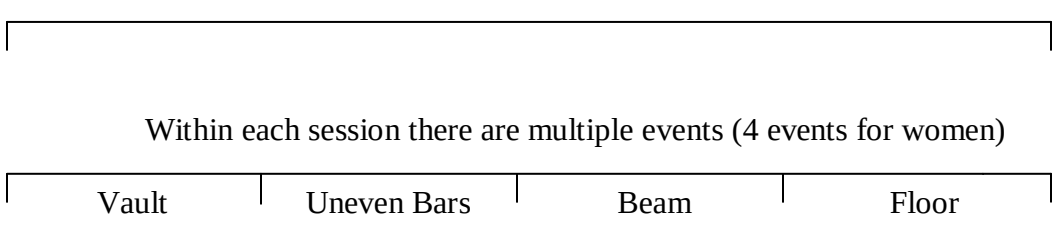
Variable	Obs	Mean	Std. Dev.	Min	Max
Order	1219	63.40689	36.59816	1	135
Order-squared	1219	5358.76	4849.31	1	18225
Normalized Difficulty Score	1219	0.000395	0.996615	-7.009	2.208469
Normalized Execution Score	1219	8.14E-05	0.996706	-9.11125	1.75576
Reputation	1219	0.053322	0.224768	0	1
Same Judge	1219	0.101723	0.302407	0	1
Male	1219	0.630845	0.482774	0	1

Performance Order

As previously mentioned, each country is randomly assigned a competition spot, which is then given to a gymnast. For example, one of the American spots was subdivision 5, starting on vault, in the fifth position. The women had five potential subdivisions during the day and the men had three. Within each subdivision the athletes started on different events; four options for the women and six options for the men.

Finally, because only one gymnast performs on the event at a time, the individual performance order was determined. Therefore, in our data athletes are assigned to a competition order on three different levels: (1) to which session, or subdivision, they will compete, (2) to which event they will start on, or their rotation, and (3) in which order they appear in their given event rotation (displayed in Figure II). Judges therefore have the opportunity to measure an athlete's performance relative to the other athletes based on the overall performance order during the entire competition, the order in which they appear in a given session, and at the smallest level, the order in which they appear in a given rotation. Throughout this study we use the overall performance order as the main control for order bias.

Figure II – Performance Order on Three Levels:



Given the previous findings, we control for the order each athlete appears in the competition and extend the literature by investigating difficulty bias. Performance quality is determined by two factors: the difficulty of the task at hand and the execution of that task. If judges are charged to evaluate the execution of a performance separate from the task's difficulty, we can determine whether task difficulty influences the execution score.

A difficulty bias exists when a participant's overall score is artificially high, or low, because of the level of difficulty attempted. This is the primary focus of this study.

Discovery of a difficulty bias in a judged event can change the optimal strategy for the participant and may lead an organizer to alter the judging process to account for, or at least test for, this bias.

Reputation

Superstar athletes are generally known in the world of gymnastics, which could create a scenario in which their reputation, or a reference bias, influences the final scores. Given previous evidence of this type of bias (Thibaut and Kelley 1959, Kingstrom and Mainstone 1985, Tversky and Kahneman 1974, Kahneman and Tversky 1996, and Findlay and Ste-Marie 2004), we attempt to reduce it by controlling for athletes who come from countries that have a reputation for producing superstars, as a proxy for reference bias. The limitations of this control are discussed in the methods section.

We define our reference proxy as those superstar countries that have won at least three medals, in the particular event of interest, in the top level competitions over the previous 9 years. This includes 3 Olympics: 2000, 2004, and 2008, as well as 6 World's competitions: 2001-2003 and 2005-2007. Superstar countries are shown in tables 4 and 5.

Table 4 – Superstar countries for women's events.

Superstar Countries (women)			
Vault	Uneven Bars	Balance Beam	Floor
USA	USA	USA	USA
Russia	Russia	Russia	Romania
China	China	Romania	
Germany		China	

Table 5 – Superstar countries for men's events.

Superstar Countries (men)					
Parallel Bars	High Bar	Rings	Floor	Vault	Pommel Horse
China	Germany	China	Canada	China	China
S. Korea	Slovakia	Bulgaria	Romania	Romania	Romania
		Italy		Poland	Japan

Country Influence

Competitions with athletes from many countries also have judges from many countries. Each event has a panel of judges designed to have a diverse set of countries represented; those judges score the same event for the whole competition. It is feared that these judges may show favoritism to athletes from their home country, resulting in a biased execution score (Zitzewitz 2006). Using data from GymnasticsResults.com, we observe the country of each judge on each execution panel.[6] We create a dummy variable controlling for whether the athlete and a judge on the execution panel in the event in which they are competing come from the same country, called Same Judge (judges' countries are presented in tables 6 and 7). Because the judges' countries are known, we do not have to worry about an anonymity bias (Zitzewitz 2010).

Table 6 – Country of the execution judges, by event.

Country of Execution Judges (women)			
Vault	Uneven Bars	Balance Beam	Floor
Mexico	N. Korea	India	Slovenia
Bulgaria	Egypt	Ireland	Germany
S. Korea	Norway	Portugal	Venezuela
Italy	Canada	Argentina	Lithuania
Romania	Brazil	France	China
Ukraine	Germany	Israel	Russia

Table 7 – Country of the judges, by event.

Country of Execution Judges (men)					
Parallel Bars	High Bar	Rings	Floor	Vault	Pommel Horse
Netherland	Algeria	Bulgaria	Japan	Mexico	Slovenia
S. Korea	Portugal	France	Venezuela	New Zealand	Russia
Lithuania	Austria	Germany	Luxemburg	Belarus	Portugal
Argentina	Ukraine	Qatar	Romania	Germany	Brazil
Czech Republic	Hungry	Jordan	Egypt	Canada	N. Korea
Poland	Great Britain	South Africa	Italy	Israel	Denmark

[6] We do not have this information for the difficulty panel.

IV. Methodology

In order to obtain an accurate measure of a judge's bias, it is necessary to separately observe two different sections of the overall score. These include the difficultly of the task at hand and the execution of the said task.

$$Score = f(Difficulty, Execution) \quad (1)$$

Therefore, the total score a gymnast receives, T, is the sum of the execution score (E) and the difficulty score (D), subtracting out any penalties (P):

$$T = E + D - P \quad (2)$$

The difficulty score is a choice variable for the gymnast, and the execution score can be thought of as:

$$E = f(O, R, J, A, D) \quad (3)$$

Where the execution score is potentially a function of performance order (O), reputation (R), country of judge (J), ability (A), and difficulty (D). It is possible that skilled judges provide a 'bonus' in the execution score when people attempt more difficult tasks. Because judges know these tasks are more difficult, they are potentially more lenient on the execution score, even when these scores should remain independent. If this is the case, that execution scores are positively correlated with difficulty, then evidence of difficulty bias exists.

Using the two different judging panels we are able to measure any impact of a difficulty bias. In order to accurately measure this bias, we control for known biases in the data: Order Bias (as shown in Flôres and Ginsburgh 1996, Bruine de Bruin 2005, and Page and Page 2010), Reference Bias (as seen in Thibaut and Kelley 1959, Kingstrom and Mainstone 1985, and Findlay and Ste-Marie 2004), and a Same Country Bias

(Zitzewitz 2006 and 2010). As a proxy for Order Bias, we include the overall performance order (O) as a measure of a given athlete's relative place in the competition and also an overall order squared term to allow for a non-linear relationship. To determine if there are a few highly talented individuals driving the results we control for a Reputation (R) as a Reference Bias. The last control captures whether a judge from a country gives athletes from their own country better scores (J). The *E* vector controls for event specific effects.[7] We also include country level fixed effects, *C*, and estimate the following for each athlete, *i*, aggregating all events, for both men and women, together:

$$ExecutionScore_i = \beta_0 + \beta_1 O_i + \beta_2 O_i^2 + \beta_3 R_i + \beta_4 J_i + \beta_5 D_i + \beta_6 D_i^2 + \delta E + \phi C + \varepsilon \quad (4)$$

To capture whether or not difficulty bias exists in the judge's decision, we add a control for the difficulty score and, to control for any non-linearietes, we include a squared difficulty score. A significant coefficient on D, difficulty score, reveals there is a difficulty bias in the judge's decision.

Recall that the difficulty and execution score, by rule, are determined by two separate panels of judges. The difficulty section scores the person for the quality of the routine, measured by how intricate and difficult the attempted skills are. The execution score is designed to measure only the execution of routine, capturing the Perfect 10 aspect that so many fans are familiar with. If difficulty and execution scores are

[7] These are set up as dummy variables for each event, women's vault excluded, and are not reported for brevity. No important results are found on the coefficients of these controls.

positively related, while controlling for the covariates described above, it will reveal biased judging, which we define as difficulty bias.[8]

Limitations

While our data are well structured for the necessary analysis to indentify difficulty bias, there are still some limitations. First, because each country's governing body places athletes into their starting positions, there are potential implications on the measurement of order bias. When countries are given starting positions, they tend to place better athletes later in the competition. This causes an upward bias in the measurement of an order bias. While we control for performance order (and therefore primacy and recency effects), an ideal dataset would have overall performance order randomly assigned to each athlete instead of each country. To our knowledge this does not exist. However, the semi-random assignment of performance order we are able to measure has no direct impact on the measurement of a difficulty bias, which is our focus.

Second, the threat of omitted variable bias presents a potential problem with accurately measuring the impact of difficulty on execution, and therefore the effect of difficulty bias. The concern is that the estimated effect of difficulty bias reflects further impacts on the execution score in addition to those from the difficulty score. In order to minimize these correlations, we control for order, same country, and reputation as described above. However, reputation may also be correlated with difficulty. For example, a gymnast with a high reputation is likely to have had high scores at previous competitions. High scores in the past are more likely if the gymnast also performed a high degree of difficulty and received high difficulty scores. Furthermore, gymnasts tend

[8] Because judges have been using the new scoring system since 2006, there has been adequate time to adjust to it. We are therefore not concerned with biases due to scoring system mistakes.

to choose similar levels of difficulty over time, creating a positive correlation between both reputation and difficulty score today. Without controlling for reputation, or reference bias as described in the psychology literature, the reputation effect may be picked up in the estimated effect of difficulty on the execution score.

We control for reputation, a country level superstar effect, but it is an imperfect measure because the rest of our data are observed at the individual level. While the gymnastics governing body (FIG) has a world ranking system based on the previous year's performance, these rankings are inefficient the year following an Olympics because there is generally high turnover in elite level gymnasts in the post Olympic year. The following year's Worlds Competition, like the one used for this study, is the coming out of the next group of elite level gymnasts and many that perform at the Olympics take the following year of competition off or retire altogether. We objectively control for reputation at the country level as described in the data section. We also test our specification with a subjective reputation measure, identifying by hand the 'big names' in the sport, and get similar results.[9] One reason for the similar outcomes may be that reputation contributes in a lesser role in judging the year after the Olympics because of the turnover. This eases any concern of a strong relationship between reputation and difficulty in our estimations. Furthermore, it solidifies that a post Olympic non-team competition is ideal for capturing difficulty bias because there is less concern with multicolinearity between reputation and difficulty with regards to Reference Bias as well as performance order and difficulty with regards to Order Bias.

Finally, there is also an issue with ability, which is unobserved but may be correlated with both the execution score as well and the difficulty score. Athletes with

[9] These results are not reported in this paper. However, results can be obtained by contacting the authors.

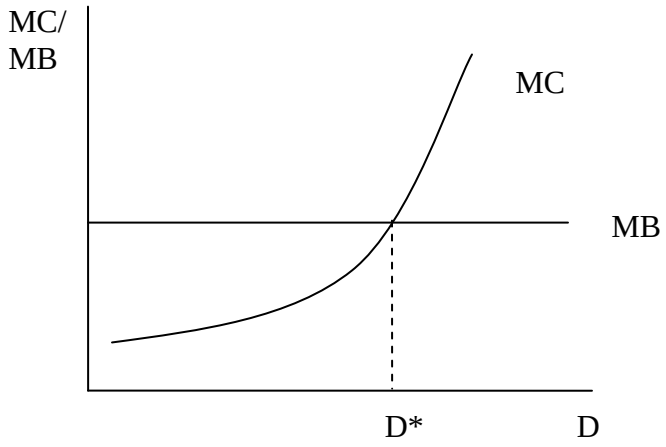
varying ability choose their own levels of difficulty, which introduces self selection concerns within our data. In an ideal situation, we would randomly assign the gymnasts different difficulty levels to measure any bias. This would presumably introduce additional variation into the execution score because some gymnasts may be asked to perform routines at a difficulty level that does not coincide with his or her optimal choice. Unfortunately this is not possible in gymnastics but it should be taken into account in other situations, such as job interview questions, where the difficulty level is determined by an outside entity.

As a gymnast chooses a difficulty level to maximize their overall score, their first order condition would be

$$\left(\frac{\partial Execution}{\partial Difficulty}\right) + 1 = 0 \qquad (5)$$

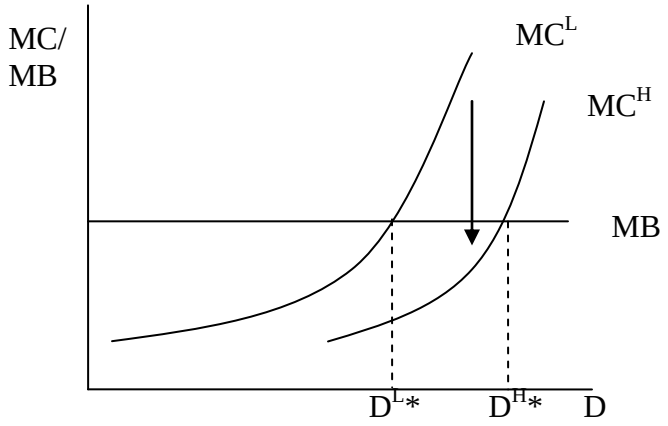
Where the Marginal Cost of increasing one's level of difficulty is $(\partial Execution/\partial Difficulty) < 0$, and 1 is the Marginal Benefit of increasing difficulty. Also, assume that $((\partial^2 Execution)/(\partial^2 Difficulty)) > 0$. Therefore, at the optimum the gymnast equates their Marginal Benefits and Marginal Costs in their choice of difficulty (D):

Figure III: The Difficulty Equilibrium



An athlete's decision on difficulty level is dependent on the sign of $((\partial \hat{} 2\ Execution)/(\partial Difficulty\ \partial Ability))$. When this is equal to zero, ability has no effect on the choice of the difficulty level. Thus, there is no correlation between ability and difficulty. However, if $((\partial^2\ Execution)/(\partial Difficulty\ \partial Ability)) < 0$, then those with higher ability levels have smaller negative impacts of attempting more difficult routines. In this case, the expected cost of a more difficult routine is lower for high ability gymnasts and they will choose a higher difficulty level. Graphically in Figure IV, a higher ability gymnast, H, will have a lower marginal cost of attempting more difficult routines relative to a lower ability gymnast, L.

Figure IV: Differing Ability Levels



Overall, the sign of $((\partial^2 Execution)/(\partial Difficulty\ \partial Ability))$ is critical to our ability to determine the existence of a difficulty bias. If this sign is negative, a control for ability is required to accurately estimate a difficulty bias. If the sign is zero, adding a control for ability does not add to the estimation's accuracy, but also does not decrease the estimation's accuracy.

Unfortunately a perfect measure of a gymnast's ability does not exist and we face an identification challenge much like the researchers attempting to capture student ability with standardized test scores or stock trader's ability with records of previous returns. We do our best to include a proxy to capture at least some of a gymnast's abilities by including country level reputation effects as described before. It is also likely that a gymnast's difficulty score captures at least part of the athlete's innate ability as well. In this study we estimate difficulty bias with and without the reputation variable and find similar outcomes.[10] We also argue that difficulty bias goes beyond acting as a proxy for ability. We encourage future research involving fine tuning the measurement of gymnastics ability.

[10] These results are available upon request.

V. Results: Difficulty Bias

To investigate whether a difficulty bias exists in the data we estimate equation 4, first without the normalized difficulty score, then including a difficulty score, and finally including the difficulty score squared term. Results of these tests can be found in the first three columns of table 8. When predicting the execution score, we find results for the existence of timing bias; competing early in the competition results in statistically lower execution scores. This supports literature finding an order bias (Flôres and Ginsburgh 1996, Bruine de Bruin 2005, and Page and Page 2010). The reference effect is positive and significant when the difficulty squared term is included; athletes from top performing countries receive higher execution scores. We do not, however, find a same judge effect. Finally, as an addition to the literature, we find a statistically significant and positive relationship between difficulty and execution scores, revealing a difficulty bias; as an athlete's difficulty level increases it artificially inflates their execution score at a decreasing rate. These results continue to hold when country level fixed effects are added in the last two columns.

Table 8 – Estimating Execution Score

Execution Score					
	(1)	(2)	(3)	(4)	(5)
O (Order)	0.008110*** (0.003)	0.008104*** (0.003)	0.008400*** (0.002)	0.007821*** (0.003)	0.008965*** (0.002)
O^2 (Order Squared)	-0.000043* (0.000)	-0.000060*** (0.000)	-0.000057*** (0.000)	-0.000052*** (0.000)	-0.000054*** (0.000)
R (Reputation)	0.730150*** (0.126)	0.046359 (0.108)	0.346858*** (0.104)	0.028198 (0.119)	0.236493** (0.113)
J (Same Judge)	0.021926 (0.093)	-0.036960 (0.077)	-0.012211 (0.072)	-0.008326 (0.076)	0.001308 (0.072)
D (Normalized Difficulty Score)		0.576618*** (0.025)	0.375128*** (0.028)	0.584243*** (0.028)	0.333762*** (0.033)
D^2 (Normalized Difficulty Score)^2			-0.121518*** (0.009)		-0.119603*** (0.010)
Constant	-0.348287*** (0.126)	-0.206036** (0.105)	-0.144924 (0.098)	-0.458479*** (0.115)	-0.279881** (0.109)
Event FE	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	Yes	Yes
Observations	1,219	1,219	1,219	1,219	1,219
R-squared	0.039	0.340	0.424	0.430	0.499
Standard errors in parentheses					
*** p<0.01, ** p<0.05, * p<0.1					

The main result from table 8, that a difficulty bias is found in the data, is economically significant as well. To put it in perspective, consider the vault score of American gymnast Rebecca Bross, who ranked twelfth on this event after the preliminary round. Bross scored a 14.250 and her difficulty score, 5.8, was one standard deviation below Un Jong Hung, the Chinese gymnast in first. If Bross attempted a one standard deviation more difficult vault, she would have not only received a .706 boost in her difficulty score, but also a .194 boost in her execution score resulting from the difficulty bias. With this bias, we estimate that a one standard deviation more difficult vault would have increased her score by .9 points, resulting in a 15.15; enough for second place. If a difficulty bias did not exist, her more difficult vault would have scored her a 14.956; placing her third. For Rebecca Bross, a one standard deviation increase in difficulty is

equal to trying the same level of difficulty as the winning athlete. On the same event, vault, the Canadian gymnast Britney Rogers scored a 14.1, with a 5.3 on her difficulty score, ranking her 15th. If she would have tried a one standard deviation more difficult vault she would have placed third, with a score of 15.0, with the difficulty bias. Without the difficulty bias she would have scored a 14.806, placing her fourth and off the podium.

It is also important to point out that attempting a one standard deviation less difficult routine has twice the impact of increasing the difficulty level. A one standard deviation increase in difficulty from the mean artificially increases the execution score by .214 standard deviations, while a one standard deviation decrease in difficulty from the mean artificially decreases the execution score by .453 standard deviations.

In addition to the difficulty coefficient being strongly significant, the R-squared is ten times higher when difficulty score is controlled for, than when it is not. This is an interesting result because when athletes attempt harder skills, it is reasonable to think they may not be able to execute as cleanly, *ceteris paribus*. Given the magnitude of the coefficient on difficulty, it is reasonable to think that the coefficient is capturing more than just a difficulty bias. We likely capture both a difficulty bias and some proxy for ability. Given this possibility, we further investigate difficulty bias and how it's related to reference bias in the next section. We also measure gender differences, judging effects, or event differences.

Interacting Difficulty and Reference Bias (Reputation)

It is possible that the known athletes are driving the results. We test this in two ways in table 9. First, we add an interaction term between the normalized difficulty score and reputation, seen in the first column. This captures the impact the reputation might

have on the difficulty score. Those athletes coming from a country with a reputation of having great gymnasts (superstar countries) receive a positive and significant difficulty bias, beyond the difficulty bias for non-superstar athletes. This shows that the difficulty bias exists and the reference bias magnifies the impact of this difficulty bias for athletes from historically successful countries. The positive and significant interaction also provides evidence that the marginal cost of attempting more difficult routine is lower for higher ability gymnasts.

Table 9 – Interaction Terms and Restricted Samples

Execution Score: Testing Reputation			
	Interaction	Excluding the Top 10%	Excluding the Bottom 10%
O (Order)	0.008949*** (0.002)	0.007437*** (0.002)	0.003780** (0.002)
O^2 (Order Squared)	-0.000053*** (0.000)	-0.000042** (0.000)	-0.000028** (0.000)
R (Reputation)	-0.251925 (0.250)	0.077316 (0.135)	0.102997 (0.082)
J (Same Judge)	-0.003502 (0.072)	-0.062450 (0.076)	0.083150 (0.056)
D (Normalized Difficulty Score)	0.316729*** (0.034)	0.237762*** (0.035)	0.213827*** (0.026)
D^2 (Normalized Difficulty Score)^2	-0.123701*** (0.010)	-0.138530*** (0.010)	0.031974* (0.017)
Normalized Difficulty Score x Reputation	0.453025** (0.207)		
Constant	-0.277673** (0.109)	-0.257633** (0.109)	-0.054104 (0.083)
Event FE	Yes	Yes	Yes
Country FE	Yes	Yes	Yes
Observations	1,219	1,095	1,099
R-squared	0.501	0.505	0.267
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

The last two columns in table 9 examine whether a restricted sample of the top or bottom 10% of execution scores are driving the results. This could potentially occur with a reference bias (reputation), because the athletes from historically successful countries

are most likely those already known by the judges. When excluding the top 10% of execution scores the results on difficulty bias continue to hold. Although this sample is smaller we find similar results, which strengthen our overall findings. This solidifies that we are not identifying a reference bias, but a separate bias towards those completing more difficult tasks.

It is also possible that the bottom 10% of the execution scores impact the results. This may occur because there is a limit of three participants per country on each event, which means a strong gymnastics nation like the United States may have to keep very talented gymnasts home, while countries not known as gymnastics powerhouses get to send athletes to compete. Because of the rule, there are contestants competing that may not have qualified otherwise. It is possible that judges award these gymnasts with higher execution scores in attempt to level the playing field with the better gymnasts. If this is the case, restricting the sample by dropping out the lower 10% would change the overall results. We find that the same judge and reputation effects are insignificant. The results for difficulty bias continue to be positive and significant, supporting the idea that judges who see a hard routine give higher execution scores when they should be independent.

In addition to the objective measure of reputation used in these regressions, we have also run all of them with a subjective measure of individual level women's superstars. We subjectively identified women's superstars by going though the data and flagging the best known names on each event; the results hold in this specification but are not presented for brevity.

Gender Differences

To measure if the difficulty bias result is being driven by the differences in men's and women's gymnastics we split the data. Scores for male gymnasts show no evidence of a timing impact, while the female gymnasts do show a timing bias; it is better for the women to go later in the competition. However, they both show strong evidence of a difficulty bias. Increasing the difficulty of a routine leads to a positive difficulty bias on the execution score, at a decreasing rate.

It is also important to note that separating male and female athletes is the only model specification that finds a same judge bias, having a judge on the panel from the country you represent matters. The same judge effect is negative and significant for the women, meaning that a given athlete is worse off when there is a judge from her country on the panel. For the men the same judge bias is positive and significant.

Judging Effects

It is possible that judges know they will be scrutinized by governing bodies of sports or researchers looking for bias. As such, judges may change their judging strategy to benefit the gymnasts they want, but in a way that is not easily detectable. For example, a judge may give a slightly higher score to an athlete from their country in the medal hunt (because it matters more for her) and give a slightly lower score to an athlete not in the hunt (because she was not going to receive a medal anyway). On average, the judge does not give a point bonus to his or her country, but they have distributed those points differently than if they had not favored their own country's athlete. This effect is measured in the interaction of the normalized difficulty score and same judge.

We do this for all athletes, as well as for male and female athletes separately. In all specifications we find an insignificant relationship for the interaction term of normalized difficulty score and same judge. This shows that when a judge is judging an athlete from their country they are not trying to hide their bias by favoring athletes that try harder routines. Male athletes continue to see a positive bias with a judge from the same country, while the female athletes have a negative impact from having a judge from the same country. We continue to find strong evidence of a difficulty bias.

Event Differences

It is possible that these results are driven by certain events, rather than gymnastics on the whole. If this is true, interacting the normalized difficulty score with each event will reveal this difference. The vault, which is a quick movement over in seconds, could yield different results than the floor routine, which lasts for a few minutes. We find no discernible pattern across events, although all events do have a positive and significant difficulty bias.[11]

VI. Conclusions

This study tests forms of judgment bias using data from elite level gymnastics. In accordance with previous literature, we control for the order of performance as well as judges from the same country and a proxy for reference bias, reputation, finding an additional form of a judgment bias: difficulty bias. In gymnastics, athletes choose the difficulty level they will attempt, introducing an issue of self selection. We also face a common identification challenge when considering a gymnast's innate ability. Despite

[11] Tables for Gender, Judging effects, and event differences have been suppressed for brevity. They are available on request to the authors.

these challenges, we find that the execution score, which is supposed to be unrelated to the difficulty score, is not; athletes who attempt more difficult routines also receive higher execution scores. This bias is magnified for athletes from well known countries, supporting additional findings of a reference bias. The reverse is also true; those that attempt less difficult routines are penalized with lower execution scores. These results hold through multiple robustness tests.

Our findings suggest an incentive misalignment for those who are being evaluated; difficulty bias may induce people to attempt more difficult tasks than they would have otherwise. The implications go beyond the world of elite level gymnastics. For example, researchers may submit more difficult projects when applying for grants, in hopes of benefitting from this new form of judgment bias. Furthermore, authors will rationally respond by including impracticably difficult statistical methods to impress referees. Musicians may optimize by choosing difficult pieces of music at auditions to impress evaluators. Employees could use unnecessarily complex presentations at work to impress a boss or gain a client.

Evaluators need to be aware of the potential issue as well, especially in situations where the participant has no say in the difficulty level. If difficulty is chosen by the judging body, a difficulty bias stresses the importance of having similar complexity for all contestants. For example, in a job interview, it is imperative that the difficulty of questions asked is similar among candidates. Otherwise a difficulty bias may exist, making it hard to accurately judge a job candidate's abilities. Political debate mediators should be aware of the potential effects as well. Candidates may benefit from being asked a difficult question during an interview or debate, even if he or she stumbles over the

response. This supports Glejser and Heyndels's (2001) idea: "it means that it is easier for an expert to compare two artists if they perform the same piece of music than if they perform different pieces", supporting the use of musical pieces with the same level of difficulty at an audition. The applications are truly endless.

This research has shown interesting insight into judging bias, with our most significant contribution as the measurement of difficulty bias. When complexity is controlled by those administering the competition, it is important that difficulty is equal amongst all candidates. When it is determined by the participant, judges should find a way to truly keep difficulty and execution separate. If they cannot, participants may efficiently respond by increasing their overall difficulty level. Continuing to search for structures that eliminate these biases as well as continued research on all forms of judgment bias are encouraged.

Works Cited

Bruine de Bruin, W., 2005. Save the Last Dance for Me: Unwanted Serial Position Effects in Jury Evaluations. *Acta Psycologica* 118, 245-260

Burgess, N. and G. Hitch, 1999. Memory for serial order: a network model of the phonological loop and its timing *Psychological Review* 106, 551–581

Campbell, Bryan and John Galbraith, 1996. Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments *The Statistician* 45(4), 521-526.

Damisch, L, T. Mussweiler, and H. Plessner, 2006. Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments *Journal of Experimental Psychology Applied* 12, 166

Emerson, John, W. Seltzer, and D. Lin, 2009. Assessing Judging Bias: An Example from the 2000 Olympic Games *The American Statistician* 63, 124-131

Flôres, Jr., R.G. and V. A. Ginsburgh, 1996. The Queen Elisabeth Musical Competition: How Fair Is the Final Ranking? *The Statistician* 45 (1): 97–104

Findlay, Leanne C. and Diane M. Ste-Marie, 2004. A Reputation Bias in Figure Skating Judging *Journal of Sport and Exercise Psychology* 26, 154-166

Garicano, Luis, Palacios-Huerta, Ignacio and Canice Prendergast, 2005. Favoritism Under Social Pressure *Review of Economics and Statistics* 87, 208-216

Gershberg, F. and A. Shimamura, 1994. Serial position effects in implicit and explicit tests of memory *Journal of Experimental Psychology: Learning, Memory and Cognition* 20, 1370–1378

Glejser, H. and B.Heyndels, 2001. Efficiency and inefficiency in the ranking in competitions: the case of the Queen Elisabeth music contest *Journal of Cultural Economics* 25, 109–129.

Goldin, Claudia and Cecilia Rouse, 2000. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians *The American Economic Review*, Vol. 90, No. 4. pp. 715-741.

Kahneman, D., and A. Tversky, 1996. On the reality of cognitive illusions *Psychological Review*, 103, 582-591.

Kingstrom, P.O., and L. E. Mainstone, 1985. An investigation of the rater-ratee acquaintance and rater bias *Academy of Management Journal***,** 28, 641-653.

Mussweiler, T., 2003. Comparison Processes in Social Judgments: Mechanisms and Consequences *Psychological Review* 110 (3), 472-489

Neilson, W., 1998 Reference Wealth Effects in Sequential Choice. *Journal of Risk and Uncertainty* 17, 27-48

Novemsky, N. and R. Dhar, 2005. Goal Fulfillment and Goal Targets in Sequential Choice *Journal of Consumer Research* 32, 396-404

Page, Lionel and Katie Page, 2010. Last shall be first: A field study of biases in sequential performance evaluation on the Idol series *Journal of Economic Behavior & Organization* 73, 186–198

Parsons, Christopher A.; Sulaeman, Johan; Yates, Michael C.; Hamermesh, Daniel S., 2011 Strike Three: Discrimination, Incentives, and Evaluation *The American Economic Review* Vol. 101, No. 4, June, pp. 1410-1435

Price, Joseph and Justin Wolfers, 2010. Racial Discrimination Among NBA Referees *Quarterly Journal of Economics* 125(4) 1859-1887, November

Rotthoff, Kurt W., (Not Finding a) Sequential Order Bias in Elite Level Gymnastics (March 7, 2013). Available at SSRN: http://ssrn.com/abstract=2230038 or http://dx.doi.org/10.2139/ssrn.2230038

Sala, Brian, Scott, John, and James Spriggs, 2007. The Cold War on Ice: Constructivism and the Politics of Olympic Skating Judging *Perspectives on Politics* 5(1), 17-29

Sarafidis, Y., 2007. What Have you Done for me Lately? Release of Information and Strategic Manipulation of Memories *The Economic Journal* 117, 307-326

Segrest Purkiss, S., P. Perrewe, T. Gillespie, B. Mayes, and G. Ferris, 2006. Implicit Sources of Bias in Employment Interview Judgments and Decisions *Organizational Behavior and Human Decision Processes* 101, 152-167

Seltzer, Richard and Wayne Glass, 1991. International Politics and Judging in Olympic Skating Events: 1968-1988 *Journal of Sports Behavior* 14, 189-200

Thibaut, J. W., and H. H. Kelley, 1959. *The Social Psychology of Groups.* New York: John Wiley & Sons.

Tversky, A. and D. Kahneman, 1974. Judgment and uncertainty: Heuristics and biases *Science* 185, 1124-1131.

Wilson, V., 1977. Objectivity and Effect of Order of Appearance in Judging of Synchronized Swimming Meets. *Perceptual and Motor Skills* 44, 295–298

Zitzewitz, Eric, 2006. Nationalism in Winter Sports Judging and its Lessons for Organizational Decision Making *Journal of Economics and Management Strategy*, Spring, 67-99

Zitzewitz, Eric, 2010. Does Transparency Really Increase Corruption? Evidence from the 'Reform' of Figure Skating Judging *Working Paper*